Algorithm-assisted decision-making in the public sector: framing the issues using administrative law rules governing discretionary power

Marion Oswald

Centre for Information Rights, Department of Law, University of Winchester
[marion.oswald@winchester.ac.uk]

Forthcoming in *Philosophical Transactions of the Royal Society A*.
Doi: 10.1098/rsta.2017.0359
Accepted: 25 June 2018

Abstract

This article considers some of the risks and challenges raised by the use of algorithm-assisted decision-making and predictive tools by the public sector. Alongside, it reviews a number of long-standing English administrative law rules designed to regulate the discretionary power of the state. The principles of administrative law are concerned with human decisions involved in the exercise of state power and discretion, thus offering a promising avenue for the regulation of the growing number of algorithm-assisted decisions within the public sector. This article attempts to re-frame key rules for the new algorithmic environment and argues that 'old' law – interpreted for a new context – can help guide lawyers, scientists and public sector practitioners alike when considering the development and deployment of new algorithmic tools.

Introduction

In 1735, in this very journal, one Reverend Barrow published a short piece, hardly a page in length, in which he surveyed births, deaths and overall population in the parish of Stoke-Damerell in Devon. [1] He notes that 'the Number of Persons who died, is one more than half the Number of Children born; and that about 1 in 54 died' in a year when the 'General Fever' infected almost all the inhabitants. He further points out that one of the persons buried was 'a Foreigner brought from on board a Dutch Ship' and two more were drowned from on board a Man of War 'but that the Ships Companies are not included in the Number of Inhabitants.' This data, together with 'Experience and Observations, both of my self and better Judges' leads him to 'reckon the Parish of Stoke-Damerell as healthful an Air as any in England.'

Fifty-four years later, we find William Morgan (communicated by a Reverend Richard Price) promoting 'the method of determining, from the real probabilities of life, the value of a contingent reversion in which three lives are involved in the survivorship.'

[2] In an age when prospects in society – and lines of credit - might be dependent on one's 'great expectations' of an inheritance, calculating the probability of achieving that inheritance (known to lawyers as contingency reversion) becomes of great interest. For instance, I might transfer a piece of land on the following basis: to my niece for her lifetime, remainder to my nephew and his heirs, but if my nephew dies in the lifetime of my niece, then the land reverts to me and my heirs; I have a 'reversionary interest' in the land. The question for my eighteenth century nephew is how to value the sum that might be payable on the contingency that he will survive his sister. The method and calculations proposed by Morgan are set out at length and in considerable detail so as to enable a reader to test and critique them. To this author's non-expert eye, two points are striking. First, that the calculations appear to be based on group data i.e. on the number of persons living at the age of my nephew, and at the end of first year, second year, third year and so, from the age of my nephew. Secondly, the article goes onto criticise a rule proposed by a certain 'Mr Simpson' and points to its results as deviating 'so widely from the **truth** as to be unfit for use' [my emphasis] in some cases producing 'absurd' results.

A modern reader might be tempted to regard these articles as illustrations of a naïve age or to a context long past, or to highlight the lack of causal evidence for Reverend Barrow's conclusion about the 'healthful' nature of his parish. Yet both articles tackle issues with which we remain concerned today: the healthiness (or otherwise) of a community, the reasons behind it and the life expectancy of an individual when compared to others. Risk forecasting and predictive techniques to aid decision-making have become commonplace in our society, not least within public services such as criminal justice, security, benefit fraud detection, health, child protection and social care. We should be better at it than our eighteenth century clergymen. It has become almost unnecessary to say that we now inhabit an information society. Information technologies driven by the flow of digital data have become pervasive and everyday, often leading to the assumption that access to vast banks of (often individualised) digital data, combined with today's networked computing power and complex algorithmic tools, will lead automatically to greater knowledge and insight, and so to better predictions.

Knowledge, however, is not the same as information (as many before me have pointed out): Knowledge, Hassan argues, 'emerges through the open and experiential and diverse (and often intuitive) working and interpreting of raw data and information.' [3] Reverend Barrow's conclusion as to the healthfulness of his parish, for instance, was based, not only on the outcome of analysis of raw data, but on additional 'experience and observations' of himself and others. Some criticise such human 'intrusion' on the data as casting further doubt on the conclusion. Grove and Meehl, a leading proponent of the use of statistical, algorithmic methods of data analysis over clinical methods, argued that 'To use the less efficient of two prediction

procedures in dealing with such matters is not only unscientific and irrational, it is unethical. To say that the clinical-statistical issue is of little importance is preposterous.' [4] It is this often-claimed superiority, together with the potential for more consistent application of relevant factors often taken from large datasets, that give algorithmic tools their appeal in many public sector contexts. Although this article is written from a legal perspective, it draws attention to arguments made in the ongoing 'algorithmic predictions versus purely human judgement' debate and applies these to the legal principles discussed below. It is particularly concerned with algorithm-assisted decisions, whereby an algorithmic output, prediction or recommendation produced by machine learning technique is incorporated into a decision-making process requiring a human to approve or apply it. 'Machine learning involves presenting the machine with example inputs of the task that we wish it to accomplish. In this way, humans train the system by providing it with data from which it will be able to learn. The algorithm makes its own decision regarding the operation to be performed to accomplish the task in question.' [5] Machine learning algorithms are 'probabilistic…their output is always changing depending on the learning basis they were given, which itself changes in step with their use.' [5] I will return to the important probabilistic characteristic of algorithmic outputs later.

Predictive algorithms and administrative law

The growth in the use of intensive computational statistics, machine-learning and algorithmic methods by the UK public sector shows no sign of abating. [6] What then should be the *role of the human* when these tools are planned and then deployed, particularly when the *accuracy* of an algorithmic prediction is claimed to be at least comparable to the accuracy of a human one? I consider this question by reference to a number of connected English administrative law rules, some of which (such as natural justice) date back to before the origins of this journal. I have done this because this body of law governs the exercise of discretionary powers and duties by state bodies, and thus the humans working within them; discretion must be exercised within boundaries or the public body is acting unlawfully. As Le Sueur explains, 'The assumption made until comparatively recently is that the decision-maker using the executive power conferred by Parliament is a human being or an institution composed of humans and that there is a human who will be accountable and responsible for the decision.' [7] We see this today in witnesses called to give evidence to Parliamentary Select Committees. The introduction of an algorithm to replace, or even only to assist, the human decision-maker represents a challenge to this assumption and thus to the rule of law, and the power of Parliament to decide upon the legal basis of decision-making by public bodies. I argue below however that English administrative law – in particular the duty to give reasons, the rules around relevant and irrelevant considerations and around fettering discretion – is flexible enough to respond to many of the challenges raised by the use of predictive machine learning algorithms, and can signpost key principles for the deployment of algorithms

within public sector settings. These principles, although derived from historic case-law, have already been applied and refined to modern government, to the development of the welfare state, privatisation, the development of executive agencies and so on.

I then attempt to re-frame each of these rules in order to suggest how they could guide future algorithm-assisted decision-making by public bodies affecting rights, expectations and interests of individuals. In doing so, I do not recommend any particular method of building or interpreting these systems [8] - as to do so would require consideration of many different contexts and informational needs - but to suggest principles to guide those engaged in future development work. I focus attention on the requirements of legitimate decision-making from the perspective of the public sector *decision-maker*, rather than from the perspective of the subject. Fair decision-making in accordance with administrative law rules by its very nature also protects the interests of the human subject of those decisions. I argue that carefully considering exactly what the algorithm is or is not predicting, and explaining to the decision-maker at the point results are displayed, is key to ensuring this fairness.

Opacity and algorithms

In contrast to the calculations of William Morgan in 1789, disclosed at length in this journal, the opacity of many of today's algorithmic tools has been much criticised, no more so than Equivant's recidivism prediction tool COMPAS. [9] Despite its deployment within the criminal justice system in the US, its workings are proprietary and remain secret.

This is not always the case however. In the UK, Durham Constabulary's Harm Assessment Risk Tool (known as HART) is one of the first algorithmic tools deployed by a UK police force in an operational capacity. It was designed to support custody officer decision-making as part of a programme called Checkpoint, which is an 'out-of-court' disposal to help a sub-set of offenders tackle their individual problems, for example drug or alcohol addiction, and so enable them to desist from crime. Durham Constabulary has been open about its development of the tool, the random forest method behind it, the input data, the first validation exercise and the challenges the tool raises in terms of officer decision-making. [10]

In the US, the Allegheny Family Screening Tool, developed by Vaithianathan and Putnam-Hornstein 'is owned by the county. Its workings are public. Its criteria are described in academic publications and picked apart by local officials. At public meetings held in downtown Pittsburgh before the system's adoption, lawyers, child advocates, parents and even former foster children asked hard questions not only of the academics but also of the county administrators who invited them.' [11]

Vaithianathan argues that transparency is never 'done': 'It starts with engaging people potentially subject to and affected by the tool, and listening and responding to their concerns. As the project continues, transparency should be revisited often to make sure that the tool is understandable to the community, agency and frontline workers.' [12] Transparency does not mean that criticism is sidestepped however [13] (and rightly so) but should mean that the resulting debate can be better informed.

Yet even if input data and algorithmic method are disclosed, 'the interplay between the two in the mechanism of the algorithm is what yields the complexity (and thus opacity).' [14] For instance, the risk prediction model created for Philadelphia's parole department by the University of Pennsylvania is made up of 500 regression trees (with an example of one of them available in a public report), the same machine learning model as the Durham tool. This report sets out the claimed benefits of this type of model as follows:

> 'The real power of random forest modeling ultimately lies in this extremely large number of separate nodes, along with the random selection of individual predictors to split them. This combination allows the influence of each predictor to be averaged over a wide variety of unique sub-samples throughout the model, and reduces the influence of any one particular tree to just one vote out of hundreds. Even if one particular branch or one entire tree proves to be somewhat inaccurate under certain conditions, therefore, its biases can easily be compensated for by the millions of other paths that cases take through the model as a whole.' [15]

This may be one of the benefits from a statistical point of view, but there is a trade-off to be had in terms of understandability, even though everything is out in the open. In the case of HART, there are over 4.2 million decision points, all of which are highly dependent on the ones that precede them in the tree structure. [10] The needs of a public sector decision-maker, or of the human subject of algorithmic decisions, are unlikely to be met by an 'information dump' into the public domain. As Guidotti et al. note, a step that is often missed is the identification of '*properties* that an explanation should guarantee.' [16] It is around these issues that 'natural justice' in administrative law can provide guidance.

Natural justice, the duty to give reasons and applicability to algorithms

Natural justice is concerned with procedural fairness, that is the control, and knowledge, of the procedure by which public bodies take action or make decisions. Although a clear understanding of the rules derived from case-law remains 'elusive', [17] it is well recognised that one of the principles of natural justice - the right to be heard - at least requires a person to be informed of the 'gist' [18] of the case against them, so that they are equipped to make representations to the decision-maker.

(These principles are reflected in Article 6 of the European Convention on Human Rights, the right to a fair trial). In relation to the position post-decision, although the courts have consistently avoided imposing any general duty in administrative law to give reasons for decisions, 'there is a strong case to be made for the giving of reasons as an essential element of administrative justice.' [19] This is particularly so where important rights or interests are concerned, such as personal liberty and where reasons would disclose a flaw in the decision-making process. The courts have also been prepared to assess the *adequacy* of reasons given, through the processes of appeal or judicial review. Sir Thomas Bingham MR asked in the case of *Clarke Holmes* whether the decision in question leaves room for '**genuine doubt**…as to what [the decision-maker] has decided and why.' [20] [my emphasis] In *Porter*, Lord Brown said in 1953:

> 'The reasons for a decision must be **intelligible** and they must be **adequate**. They must enable the reader to understand **why** the matter was decided as it was and **what conclusions** were reached on the 'principal important controversial issues'…The reasoning must not give rise to a **substantial doubt** as to whether the decision-maker erred in law, for example by misunderstanding some relevant policy or some other important matter or **by failing to reach a rational decision on relevant grounds**.' [21] [my emphasis]

The incorporation of an algorithm into a decision-making process may come with the risk of creating 'substantial' or 'genuine' doubt as to why decisions were made and what conclusions were reached, both for the subject of the decision and the decision-maker themselves. The use of an algorithm is not likely to provide an excuse or justification for a lesser standard: Lord Carnwath stated in *Dover District Council* that 'the content of [the duty to give reasons] should not in principle turn on differences in the procedures by which it is arrived at.' [22] It could certainly be argued that, as the *Doody* decision held that a Home Secretary setting a tariff of imprisonment must show 'how his mind is working' [23], the same will be true of a human taking an algorithmically-informed decision. It could even be said that the use of an opaque algorithm to generate an output informing a decision might obstruct the right to be heard if an individual is unable to understand the 'gist' of why the output was generated and so present an alternative case, or if the 'learning' nature of the tool makes it difficult or impossible to recreate the original decision.

Hildebrandt is concerned that the provision of information or explanations should not be mistaken for legal justification of a decision. [24] She states as an illustration: 'When a court decides a case it cannot justify its decision by spelling out the heuristics of the judge(s) involved, such as their political preferences, what they had for breakfast or how they prepared the case.' [24] While it is certainly the case that explanation does not necessarily imply justification, a duty to give reasons so as to avoid 'substantial doubt' can reveal flaws in the process, the sort of error which

would allow the courts to intervene, and provide information required for audit and thus justification, or otherwise. This should identify whether there was a disagreement between the human decision-maker and the algorithmic recommendation or prediction and if so, why the algorithmic recommendation was followed (if it was). Evidence of 'meaningful' human involvement will be vital to demonstrate that a decision was *not* automated processing as defined by EU data protection law. [25] Also, setting out the explanations for an individual prediction could reveal errors that have legal consequences, for instance leading to a public body acting outside its powers or unfairly. At an earlier stage, a duty to provide the subject with the 'gist' of the factors weighing against them (in an algorithmic risk assessment for instance, the most important factors that informed the risk assessment) could enable the individual to argue the alternative during the process. [26]

A higher standard for algorithms?

In applying these requirements to algorithm-assisted decisions however, would we be unfairly requiring a higher standard of algorithms than we are of humans? It is indeed the case that in many walks of life when we interact with the public sector, we defer to recommendations of humans without always requiring a detailed breakdown of why they have reached a particular recommendation, in the medical field for instance. Yet those decisions are neither unjustifiable nor unexplainable. The medic carries out her work within a regulated structure that involves training, certification and ongoing oversight, subject to a legal framework that allows decisions and actions to be challenged. The fundamental principles around patient informed consent require a medic, *inter alia*, to inform the patient about the diagnosis, including any uncertainties: 'If you recommend a particular treatment or course of action, you should explain your reasons for doing so.' [27] There may be reasons for not sharing information with patients (for instance around capacity, the risk of causing serious harm or patient wishes), but the medic must be prepared to explain and justify her decision not to share [27]. In any event, we would expect her to have sound reasons for the diagnosis and treatment recommendations, even if those are not explained in detail to the patient.

Judges operate within established frameworks, in England & Wales one that regards judicial competence, independence and accountability - through appeal and scrutiny - as of crucial importance (and not just in England and Wales!). [28] The famous study on judicial decision-making before lunch is often cited as evidence of the frailty of human judicial decisions and the influence of hidden factors, [29] although various criticisms of this study include one which focuses upon the overlooked factor of decision pattern, concluding that 'the phenomenon of favourable decisions peaking after a meal break is likely an artefact of the order of case presentation.' [30] Another concludes that the same effect could be produced by rational time-

management factors. [31] Pasquale and Cashwell dismiss the assertion that judicial opinions are more opaque than machine learning algorithms: 'Unlike many proprietary or hopelessly opaque computational processes proposed to replace them, judges and clerks can be questioned and rebuked for discriminatory behaviour.' [32]

There remains the risk of course that post-event explanations or justifications of a human decision only partially represent the 'real' reasons. But administrative law principles governing the way that state actors take decisions via human decision-makers, combined with judicial review actions, evidential processes and the adversarial legal system are designed to counter this sort of practice. The incorporation of the outputs of algorithmic tools, that may represent a 'digital unconscious' as Hildebrandt has put it [33], into a decision-making process will not be exempt from this oversight.

Some forms of automation hold out promise for legal certainty - 'like cases are treated identically, the elimination of bias, ensuring that no irrelevant considerations are taken into account, and that all relevant factors are included.' [7] Advances in medical AI could revolutionise the ability for medics to detect a patient's long term trend 'without wading into the data themselves.' [34] It does not follow from this, however, that there should be less provision made for understanding and querying outputs. [7] Requiring algorithmic systems to provide explanations for their recommendations, suitable to each particular context, would make a positive contribution to the rule of law. [7] Polson and Scott argue that one of the advantages of algorithms is that the biases of 'human wetware' i.e. the human brain, cannot be subjected to direct numerical scrutiny in the same way as algorithms, although they regard the secrecy around the COMPAS algorithm as 'morally obscene.' [34] I would argue that it is not a *higher* standard that would be required by administrative law principles for algorithm-assisted decisions, but one that is adapted to the way that an algorithm-assisted decision is structured.

Algorithms and intelligibility

It would be a mistake to regard the law as disconnected from the aims and objectives of public authorities, somehow operating in a vacuum. The system of administrative law is not a barrier or 'antagonistic' to efficient government it is a 'creative' not destructive relationship, focused on improving the 'technique' of government, and thus the confidence of the citizen in its reasonableness and fairness. [19] Developments in algorithmic intelligibility and explainability can improve 'techniques' of government, and administrative law principles can inform the requirements for such intelligibility, an approach with fairness as its goal. This goal seems in harmony with the aims of many in the data science field working on explanations for data-

driven classifications. Martens and Provost, in their article on explaining data-driven document classifications, argue that:

> 'We need research that focuses on a user-centric theoretical understanding of the production of explanations with a primary goal of improving data-driven models based on feedback and iterative development. This is important because as model-based systems increasingly are built by mining models from large data, users may have much less confidence in the model's reasoning than with hand-crafted knowledge-based systems. There are likely to be many cases where the decisions are erroneous due either to biases in the process, or to over-fitting the training data.' [35]

Rather than a 'passive recipient of explanations about why she is wrong about the world', a user would see herself as an active part of the system development. Martens and Provost also identify the need to differentiate between the different roles of people interacting with the system in terms of the explanations provided, for instance the manager who may need to sign-off models or explain or justify models in the case of error. [35] (Such 'analytical quality assurance' established by the senior accountable person is a requirement of HM Treasury's 'Aqua Book', the guidance on producing quality analysis for government. [36])

Martens and Provost are particularly concerned with explanations that can have an impact on improving the model as well as improving user acceptance. What should the role of the human be, however, in circumstances when studies are said to show the superiority of statistical prediction? [37] [38] Berk and Bleich comment that 'one does not have to understand the future to forecast it with useful accuracy…Understanding a phenomena may lead to improved forecasting accuracy, or it may not, but forecasting and explanation are different enterprises that can work at cross-purposes.' [39] They question what a judge would do with an explanation as to why an individual was forecasted high or low risk. [39]

Referring back to the principles of natural justice discussed above, one of the fundamental reasons why a judge (or other public sector decision-maker) would need an explanation regarding an algorithmic forecast is to determine whether or not there was a flaw in the overall decision-making process that had been informed by the algorithm, or indeed whether her own decision risks being affected by such a flaw. In relation to pre-emptive policing, Hildebrandt warns 'those meant to be pre-empted are left in the dark, while those employing the predictive analytics have a hold on the steering wheel (though they are probably far less in control than they may be inclined to believe)' [33] Operating an algorithm-assisted process in accordance with administrative law principles may enable the public sector decision-maker to keep control of the algorithmic 'steering wheel' and operate it in a lawful manner.

It is accepted that there is evidence to demonstrate that '**When the data bases are identical**, the findings have been uniform in showing that statistical *combination* of data is superior to clinical combination.' [my emphasis] [37] The responsible decision-maker must be the one to determine, however, whether the database used by the algorithm is indeed identical i.e. that it represents all the factors that should be taken into account (more on this below). In addition, the human must determine whether the decision under consideration matches the one for which the algorithm was developed – for instance, an assessment of 'risk' may encompass much more than the forecast of a particular behaviour by an algorithm - and whether the data on which the algorithm was trained match the circumstances of the current situation. Polson and Scott comment 'A machine can make predictions based on the assumptions with which it's programmed, but only people can check those assumptions.' [34] Van Kleek et al. refer to algorithmic explanations as 'sensemaking' and explain the need for this as follows:

> 'In order to grapple with value-laden decision-making, practitioners both individually and collectively need to build up mental models of the decision support systems that they work with. These schemas allow for more nuanced evidence than the raw yes/no output from the algorithm. In cases where algorithms are being used to make more open ended decisions, being able to spot situations where algorithmic output is expected to be flawed is of real value in shaping questions and further examination. An example of this might be a predictive policing system exaggerating the risk of crime in an area immediately following a festival or carnival. If operators understand (or infer) that the system makes use of incident rates from previous months, then they can interpret, or even predict, its output accordingly.' [40]

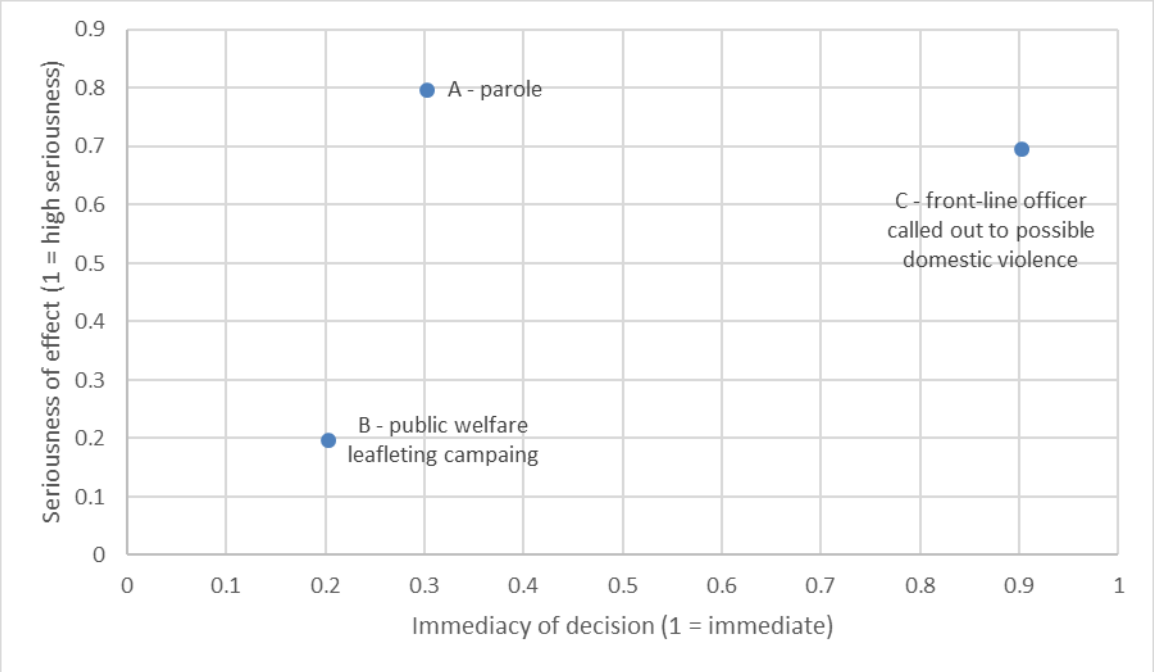
Finally, as Barnes comments, 'what are you going to do with the [predicted outcome]?...Once you have a red box on your computer screen that says high risk, now what?' [41] Appropriate knowledge as to why a prediction was generated will be necessary in order to decide upon the circumstances in which a prediction should be listened to, queried or overridden. Furthermore, there is a logical step required between presenting a prediction and interpreting it as a risk.

Properties of an explanation

This is not to say that everyone involved in all stages of a decision-making process will require the same explanation regarding an algorithmic output. There will be different '*properties'* or granularity that should be provided by an explanation dependent upon the context, the particular user and the likely weight of the outcome that the algorithmic output informs. For instance in a policing context, management staff (responsible for ensuring that the tool is fit for purpose and for monitoring the legality of the force's actions) are likely to require more information about the long

term performance and accuracy of a tool when compared to a front-line officer dialling up a risk predictor in a confrontational operational situation.
Putting aside whether there should always be some sort of base-line standard, Figure 1 below attempts to illustrate the potential bearing on the granularity of reasons required of *immediacy* of decision and *seriousness* of outcome or effect (that is, the weight of the impact on rights or freedoms of individuals).

Fig. 1



In situation C (a police officer called out to a possible domestic violence situation), the officer may be required to decide rapidly as to whether or not to make an arrest. Providing the officer with details about the uncertainties of the output of any algorithmic decision-support tool or highlighting any borderline risk assessment may not contribute to the most effective decision-making process in those circumstances. Yet a decision either way informed by unreliable information could have significant consequences for both the offender and victim, suggesting that such details should be available to others within the organisation with oversight and management roles to ensure accountability. In situation A, a decision about an individual's application for parole can be taken in longer time. As we have seen recently however from the case of convicted rapist John Worboys, it is one that can seriously infringe both a victim's and offender's rights and freedoms [42] and therefore the standard for algorithmic intelligibility should be commensurate with that high potential impact. Situation B represents the sort of public sector decision that could be informed by algorithms but which has a low immediacy and a low seriousness, in terms of the impact of the decision, and where there may be limited external factors to consider. The principles of natural justice and the duty to give reasons may have little applicability to such circumstances.

To conclude this section, in each algorithmic-assisted environment, a context-specific and nuanced approach will be required so that the information and explanations provided to aid intelligibility, or the way the result is interpreted, enable the particular public task to be fulfilled in a legitimate manner. The design of interpretable tools should take into account both the requirements of natural justice and the practicalities of 'the messy, socio-technical contexts in which they inevitably exist.' [43] As Christin has argued, we need to pay attention 'to the actual rather than *aspirational* practices connected to algorithms.' [44]

Duty to give reasons – reframing for algorithm-assisted decision-making in the public sector

The first of my suggested 're-framings' focuses upon key factors within the right to be heard and duty to give reasons in order to suggest how they could guide future algorithm-assisted decision-making by public bodies:

When an algorithmically generated prediction, recommendation or other output forms part of a decision-making process, consideration should be given to the circumstances in which reasons for/an explanation of the output may be required. These may include, *inter alia*: to determine whether the data on which the algorithm was trained match the circumstances of the current situation; the identification of situations where the output is likely to be flawed; where individual rights and freedoms are under consideration. The properties or granularity that should be provided by an explanation will be dependent upon the context, the particular user requiring the explanation and the likely weight of the outcome that the algorithmic output informs.

The next section will discuss irrelevant and relevant considerations. Linked with the above, only if we know the grounds on which a decision has been taken, can we judge their relevance. Reverend Barrow's opinion as to the healthful nature of his parish was based upon data, together with 'Experience and Observations, both of my self and better Judges.' [1] We have no further knowledge, however, of those experiences and observations, nor of the grounds on which they were made, nor of their weighting compared to the data, and therefore can make no meaningful assessment of them. The grounds on which algorithmic predictions or recommendations are generated are commonly just as obscure, if not more so, than those relied upon by Reverend Barrow. If such prediction or recommendation forms an important element of the decision-making process, how then can its lawfulness in terms of relevance or irrelevance be judged?

Irrelevant and relevant considerations

Many administrative law cases are concerned with a public body's alleged improper motives or where it has acted upon irrelevant considerations. The doctrine was explained by Lord Esher MR as follows:

> 'If people who have to exercise a public duty by exercising their discretion take into account matters which the courts consider not to be proper for the exercise of their discretion, then in the eye of the law they have not exercised their discretion.' [45]

One of the most well-known cases is *Venables* where the Secretary of State had fixed a tariff for two boy murderers. It was held that he had misdirected himself in taking account of public demands and newspaper campaigns when coming to his decision. [46]

Relevancy and algorithms – data inputs/predictors

Different definitions of 'relevance' are potentially in play. Berk argues 'if other things equal, shoe size is a useful predictor of recidivism, then it can be included as a predictor. Why shoe size matters is immaterial.' [39] In this statement (I imagine made somewhat in jest!), we find a potentially significant area of friction between lawyers and data scientists. Lawyers are likely to question whether shoe size would be a relevant consideration if a risk assessment was made via other means? If it would not, should it be an input factor in an algorithmic assisted prediction? Maybe though, lawyers are talking at cross-purposes with statistical experts. Administrative law is concerned with situations where a public authority has acted upon irrelevant considerations or failed to take into account relevant ones, in relation to the power or duty being exercised, and so its action or decision is *ultra vires*, beyond its powers and therefore null and void. Those working on algorithmic risk prediction tools are particularly concerned with the relevance – the statistical correlation - of a factor to model's predictive performance Indeed Barnes and Hyatt comment 'since there is little penalty for including additional predictors – even when they add little in the way of predictive power – a wide variety of different predictors can be used to construct these models.' [15]

Where there is only correlation between the factor and the output, and limited causal evidence, will the use of this factor be defensible if the output informs a decision by a public body, especially if removal of the factor affects accuracy of the tool? Accuracy may though not be the overriding concern for all public bodies. Multi-layered public policy considerations, such as community and social engagement, might outweigh the arguments in favour of including a factor, even if removing that factor reduces the accuracy of the tool. Other legal duties imposed on public authorities, for example the Public Sector Equality Duty under the Equality Act 2010, will require a more holistic consideration of the impact of an algorithmic tool and how it might affect different groups in different ways. In addition, the quality of the factor itself will be

significant to its defensibility. We would all look aghast nowadays at the 1926 Court of Appeal judgment that upheld a decision to dismiss all married women teachers, on the basis that the discretion of the public authority in relation to the efficient maintenance of their schools had not been exceeded. [47] How would we react however to inclusion of marriage as a predictor, perhaps not too far away from demographic predictors [10][15] or census poverty indicators [48] used in recent criminal justice models? The developers of these particular tools are likely to argue that these factors are related to the purposes of the programmes of which the algorithmic tools are part (in the case of Durham's HART tool for instance, to tackle the cycle of repeat offending in certain communities), even if direct evidence of causation may be lacking. The courts have been prepared to allow public bodies a relatively wide discretion to take into account a range of legitimate factors in their decision-making and so may be reluctant to uphold a challenge *in administrative law* to the use of predictors – of itself - that have an explainable (in terms of some degree of causation), non-biased and potentially justifiable link to the purpose in hand. Use of shoe size is likely to be another matter!

Relevancy and algorithms – outputs

As further evidence becomes available in different environments as to algorithms' accuracy in comparison with the human decision-maker, we may start to hear arguments that outputs should be regarded as relevant considerations, from a legal perspective, and indeed that public authorities should proactively seek out such algorithmic outputs. The court in the Worboys challenge said that the parole board should have sought out further information as to the circumstances of his offending. [42] Perhaps in the future such further information will always be expected to include an algorithmically generated risk assessment.

A few notes of caution however: a determination of relevancy could stand or fall on the tool's performance in the live environment, the relative importance of extrinsic factors (discussed in the next section) or whether the model is predicting the right thing. In terms of performance, a model that was trained on adult male offenders' data for instance may have little relevance to female or juvenile offenders. An output could be undermined, Cabitza et al. have argued, if inputs are based only on the values that have been proposed by a statistically significant majority in order to sweep 'uncertainty under the carpet.' [49] Furthermore, a public body's remit is often dependent on a subjective assessment (say 'in the reasonable opinion' of an official). Take a hypothetical example of a government department which has been given a statutory power to intervene 'if a child is reasonably determined to be at high risk of harm.' The legality of its power to intervene is therefore dependent upon this assessment. The government body might use an algorithm to help it decide on risk levels by way of textual analysis, of hospital admission reports for instance. If it intervenes in respect of a child at low risk, and if this assessment was unreasonable

due to issues with the tool, then it will be acting outside its powers (not to mention reducing services to children at real risk of harm and causing unnecessary disruption and stress to the child and family).

In this context, the work of Ribeiro, Singh and Guestrin is of interest. Their experiment - in respect of classifiers that were trying to determine if a document was about Christianity or Atheism - used a dataset that, despite a high accuracy on validation data, contained features that did not generalise, and thus validation accuracy overestimated real world performance. [50] When test set *accuracy* was used as a measure of trust in order to choose between two text classification models, users tended to select the worse classifier i.e. the classifier that despite achieving a high percentage accuracy rate, in fact had serious issues 'in the wild' due to the importance given to irrelevant words (such as 'Posting' and 'Re'). When individual prediction explanations were shown, however, it was possible for a human user with prior knowledge to see if a prediction was made for arbitrary reasons unconnected with the purpose of the prediction, and so take steps to improve an untrustworthy classifier. [50]

Risk assessments and predictions

Riberio et al.'s experiment was in respect of a definitional classification. Risk assessments present more of a challenge to human judgement where, Meehl would have argued, human judgement does not represent a gold standard. Non-expert users may struggle to understand the factors that contribute to the algorithm's output, such as the 'not-so-obvious words' that contributed to the classification of pornographic websites in Martens and Provost's work, [35] and therefore could be set up to fail if asked to decide upon relevance.

This however brings us back to the potential disconnect regarding the meaning of 'relevance'. In a public sector environment, relevance in a legal sense cannot be ignored and so neither can statistical relevance - and the risk of false generalisation - to this assessment. Many lawyers will find themselves in the position of having to consider whether the results of a model satisfy a certain relevancy or evidentiary standard or whether use of a model with say, an 65% overall accuracy rate, satisfies a certain standard of care. How should they decide whether statistical algorithmic risk predictors are 'unreliable science' [51] or in fact empirically valid in any one context?

Their difficulties are only exacerbated by the 'group-to-individual' problem. Melissa Hamilton argues that translating from the population, being the group level, to the individual level 'is a precarious adventure fraught with errors; but many judges, practitioners, even forensic assessors, fail to notice.' [51] She points to misleading communications during sentencing decisions in the US in which group-based data was translated into absolute predictions of reoffending at an individual level. Therefore, Hamilton argues, attention must be paid to the way that results of algorithmic tools

are communicated to decision-makers, with Hamilton advocating a comparative or analogous form of risk communication.

Blastland and Spiegelhalter describe the challenge another way: 'the average can be scarily predictable, but only at the right scale. This is the scale of whole populations, boiled down and their essence extracted. The problem is that this is not the scale on which individuals in all their variability live.' [52] Even Richard Berk has described his tool's forecasted low risk offenders as 'good bets'. [39] In presentation of algorithmic results to the human user in practice, however, it may be less than clear that, as the Supreme Court of Wisconsin pointed out in the *Loomis* case: 'risk scores are intended to predict the general likelihood that those with a similar history of offending are either less likely or more likely to commit another crime following release from custody…the risk assessment does not predict the specific likelihood that an individual offender will reoffend. Instead, it provides a prediction based on a comparison of information about the individual to a similar data group.' [53]

Pasquale and Cashwell point to the paramount importance of 'meaning' in rights determination, not factored into many predictive models. [32] Indeed, there is increasing political pressure to consider 'how unjustified correlations can be avoided when more meaningful causal relationships should be discernible' with transparency proposed as a default when the algorithms in question affect the public. [54] In terms of design solutions to these issues, counterfactual methods [26] and causal reasoning [55] have the potential to provide users with the information that they need to consider the defensibility of the output while not requiring expert knowledge around the statistical relevance of the input factors, provided that this information is provided at the point of result publication, not just in training.

This is not to say that outputs of predictive algorithms automatically cross the line into irrelevance: risk assessment is an essential part of many public services. Hofman, Sharma and Watts argue that 'social scientists could benefit by paying more attention to predictive accuracy as a measure of explanatory power.' [56] A decision as to whether to refer someone onto a deferred prosecution scheme, such as that offered by Durham Constabulary, inevitably involves an assessment of risk - of 'future dangerousness' as Richard Berk would say – as well as a consideration of personal circumstances, something that could easily be a rather cloudy, hunch-based decision, not to mention a difficult one. Could algorithmic risk assessment tools, as Blastland and Spiegelhalter put it, 'work well enough to give…a practical steer'? [52] Should it matter if the causal relationship is uncertain if a model is designed for circumstances when its ability to predict may be one of the important factors?

Predictive accuracy and extrinsic factors

Hofman et al. point out the need to assess whether 'predictive accuracy is subject to some fundamental limit' because of dependence upon extrinsic random factors. [56] Such limits must be considered before determinations about an algorithm's relevance, or otherwise, can be made. [57] For instance, the national Domestic Violence Disclosure Scheme in the UK, known as Clare's Law (named after Clare Wood who was murdered by her ex-boyfriend who had a history of violence towards women), has as its very heart an assessment of risk by the police, requiring them to make a judgement as to whether to disclose information about an individual to the person with whom they are forming a relationship. [58] They can only do this if there is a 'pressing need' for that disclosure i.e. that the risk of harm reaches a certain level. There is potential for algorithmic tools to help officers to make 'better bets' or even improve the currently rather opaque decision-making process relating to this information sharing scheme. The debate around the use of actuarial algorithmic tools could present an opportunity to clarify the sort of risks that would result in a 'pressing need' for disclosure, and what factors should go into that assessment. Of course, Clare's law, and domestic violence in general, is a factually and emotionally complex sphere requiring a focus upon the (potential) victim as well as the perpetrator, and an appreciation that stark conviction and arrest data will not often present the full picture.

Irrelevant considerations – reframing for algorithm-assisted decision-making in the public sector

Eighteenth century Morgan would have appreciated this search for relevancy; his criticism of the pre-existing rule was based on its absurd results in the real world, its lack of 'truth', such as the probability of an 18 year old surviving one 78 year old being calculated as less than the 18 year old surviving two 78 year olds. [2] The pre-existing rule neither reflected observation nor common sense. As Mulgan points out:

> 'Everything we know is knowledge from the past, which may not apply in the future – the problem repeatedly stumbled on by models, algorithms, economic theories, and geopolitical dispositions, which made sense in one era, but then become dysfunctional in another….And so the models we use to think can also become traps…intelligence has to be at war with and suspicious of itself to be truly intelligent.' [59]

But relevance is not an easy concept to define – it means different things to different people - and there seems to be much work to do to achieve understanding between the various disciplines involved in the creation, deployment and regulation of algorithms, and in particular to determine the defensibility of predictors which are key to predictive accuracy: 'only people can decide which data points are appropriate to use in the first place.' [34]

The second of my ‘re-framings’ does not attempt to solve this dilemma but instead to set out a number of factors that may need to be considered in an assessment of legal relevancy when an algorithm-assisted decision is in play:

In deciding upon the relevance of an algorithmic output to a decision by a public sector body, the human decision-maker should consider *inter alia* a) the **relevance of the input** factors to the context of the decision, in particular whether they have an explainable and ultimately justifiable link to the purpose in hand; b) the tool's performance and accuracy in the **live environment**; c) the relative importance of **extrinsic external factors** (those not factored into the algorithm) to the overall decision**;** and d) the level of uncertainty around **causal** relationships between the inputs and the prediction claimed.

Improper delegation and fettering discretion

Discretionary power is crucial for effective government: 'Relatively little can be done merely by passing Acts of Parliament. There are far too many problems of detail, and far too many matters that cannot be decided in advance.' [19] Discretionary power must be not abused, either by 'running amok' or by failing to exercise discretion appropriately. There is a significant body of case-law dealing with improper delegation of powers allocated to a public body and ‘fettering’ of an organisation’s discretion when exercising powers. Often Parliament has created a scheme, whereby it clearly intended that a particular person or body should make the decision in question, not the person to whom the discretion has been delegated. [60] Regarding delegation, Wade & Forsyth explains that ‘the vital question in most cases is whether the statutory discretion remains in the hands of the proper authority, or whether some other person purports to exercise it.’ [19] Improper delegation might include putting a decision ‘into the hands of a third person or body not possessed of statutory or constitutional authority’ [61] or abdicating powers, such as where the Home Secretary acted as a ‘rubber stamp’ on the advice of others without making his own decision. [62] Where a public sector body is given an element of discretion, it must put its mind to the decision and not follow policy or other diktat blindly. A general policy or rule is acceptable provided that, as Lord Reid said in *British Oxygen,* the authority does not refuse to listen at all. [63] An administrative authority is not allowed to ‘pursue consistency at the expense of the merits of individual cases.’ [64] Hildebrandt sums up the importance of discretion to the application of decisions in the public sector affecting individual rights: ‘Discretion, rather than strict application of unbending rules, recognizes the fallibility of interfacing rules with their field of application.’ [33]

Fettering discretion and algorithms

A public body whose staff come to rely *unthinkingly* upon an algorithmic result in the exercise of discretionary power could be illegally 'fettering its discretion' to an internal 'home-grown' algorithm, or be regarded as delegating decision-making illegally to an externally developed or externally run algorithm, or having pre-determined its decision by surrendering its judgement. Hildebrandt has used the term 'judgmental atrophy' to describe these outcomes [65] and notes that data-driven architectures can 'transform the environment we depend upon, while also transforming ourselves in the process.' [24] For the public sector, this must include the risk that the deployment of an algorithmic tool starts to change, or limit, the way that a decision is taken or an operation is carried out. Karen Yeung describes being 'hypernudged' in one direction, [66] a risk when an algorithmic output is expressed in very blunt un-nuanced terms (high, low and so on), and where the algorithm's workings are opaque to the human user, thus bringing us back to the first principle discussed above. Gary Kasparov says that 'The problem comes when the database and the engine go from coach to oracle'. [67] This reflects problems from an administrative law perspective which come when 'the real discretion is being exercised by the body or person that recommends.' [19] Or in the algorithmic examples that we are considering, where the discretion has been delegated to the algorithm, and no genuine or conscious choice is being made by the public authority.

The role of the human in an algorithm-assisted discretionary decision?

An algorithm has the potential always to be more accurate than a human when circumstances are identical to that which the algorithm was developed, and in relation to that specific task or question for which the algorithm was designed (although accuracy levels may not present the full picture in circumstances where data uncertainty is not represented [49] or where there is no appetite for double-blind testing, in respect of serious offenders for instance: see Kleinberg et al.'s work on comparing judicial bail decisions with an algorithmic one where crime outcomes can only be observed for released defendants [68]). Grove and Meehl argue that 'if an equation predicts that Jones will do well in dental school, and the dean's committee, looking at the same set of facts, predicts that Jones will do poorly, it would be absurd to say, "The methods don't compete, we use both of them." One cannot decide both to admit and to reject the applicant; one is forced by the pragmatic context to do one or the other.' [4] It is hard to take issue with this. It is the most accurate prediction that should be deployed, Grove and Meehl say, which may in many scenarios be the algorithmic one. Where there were exceptions to the superiority of algorithmic assessment in various studies, it was argued that had the data that was available to the clinician been made available to the predictor, then the statistical predictor might have been equally or more accurate; the solution was to improve the model, rather than to combine human heads with the model. [4] Humans struggle to correct patterns of ineffective or biased decisions, often due to

the lack of meaningful feedback, whereas algorithms are specifically designed to learn through error. [4]

Algorithms deployed in public sector environments will inevitably be limited in their data inputs, for legal, technical and policy reasons too numerous to explore here. For instance, the HART tool deployed by Durham Constabulary to forecast risk of serious offending in the context of triaging offenders to an out-of-court disposal currently uses only data available in local constabulary systems, not data from neighbouring forces or available in other systems such as the Police National Computer. In such circumstances, it is therefore the human decision-maker that will have the knowledge of factors that are not represented by inputs in the algorithm, including 'procedural and tacit' knowledge acquired from hands-on experiences and practice [69], or employment and family circumstances (as Kleinberg et al. acknowledge in their study [68]), social contacts, disability or mental health, the positive affect of interventions, the circumstances of the victim or community or intelligence records, which commonly require human interpretation, perhaps of a link with organised crime not reflected in arrest, charge or conviction history. They must be allowed to, and must be expected to, take all relevant factors into account, and to record how they do this.

Improving a model with new factors

But could not such factors be added to a model in order to improve it? Indeed they could, and in doing so the creators of algorithms inevitably need to 'translate 'real' life events into machine readable data and programs', [24] with potential consequences for the exercise of discretion. First, factors such as family and social relationships, or impact of health conditions, are not easy to 'datafy' although the COMPAS tool appears to attempt to do so, based on the questions in the questionnaire presented to offenders. [70] Hildebrandt states that 'as with every translation, something gets lost' [24] and there must be a risk that 'datafication' of such factors might change them into too simplistic a format. Even if it appears that all relevant inputs have been captured, the human decision-maker must be alive to the relevance of those 'lost' elements. Otherwise, it could be tantamount to the decision-making being brought forward to the technical stages of a system's design. [5] Furthermore, someone will always present with 'factors' that are relevant but for which the algorithm was not trained. The human decision-maker must not refuse to exercise their discretion to consider such factors.

Secondly, we cannot always assume that the forecast or classification represents the only or main factor on which the 'rightness' or 'wrongness' of the overall decision is to be judged. Doing so may risk changing the question that the public sector decision-maker has to answer. *Young Jones was admitted to dental school despite the algorithmic prediction that he would do poorly, and look he has done poorly, therefore the human decision was wrong*. But perhaps the University's policy of

admitting candidates from deprived backgrounds outweighed the prediction at the time. *The offender was predicted to be medium risk by the algorithmic tool, but the police officer decided to release her on bail, and look she has reoffended, so the human decision was wrong*. But at the time of the decision, the offender was assessed to have strong family and community ties with a plan to address her drug problem. The College of Policing Authorised Professional Practice's risk principles state, 'By definition, [operational] decisions involve uncertainty, ie, the likelihood and impact of possible outcomes cannot be totally predicted, and no particular outcome can be guaranteed.' The principles go on to say 'assessments of decisions should concentrate on whether they were reasonable and appropriate for the circumstances existing at the time. If they were, the decision maker should not be blamed for a poor outcome.' [71]

The question to be answered

Polson and Scott point out that 'A machine can fit a model, but only people can use that model to ask the right questions.' [34] Questions and decisions based on risk, and legal concepts such as 'reasonableness', 'public interest' and opinions of necessity represent a challenge for algorithms and for feature engineering [34]: to produce a model that is genuinely able to reflect the complexity of individual circumstances, which apply to the multiple elements that may need to be considered, and which produce every choice of next steps that could reasonably apply to the decision(s) in question. The UK's Police and Criminal Evidence Act, for instance, states that a person shall be released after charge with or without bail unless the custody officer has 'reasonable grounds' for believing that the detention of the person arrested is necessary to prevent him from committing an offence… to prevent him from causing physical injury to any other person or from causing loss of or damage to property… to prevent him from interfering with the administration of justice.' [72] Although risk of offending will be a relevant factor in this decision, it does not represent the question to be answered, which is whether there are *reasonable grounds* for believing that detention is *necessary to prevent* the offender from committing an offence and so on.

Furthermore, as discussed above, the forecasts produced by many existing algorithmic tools are probabilities (that the person or situation in question has a certain similarity to people or situations in the past). But they appear at times to be presented as something more: a prediction of reoffending becomes a 'risk' of reoffending and thus the risk if, say, a person is given parole. Determinations of risk – a decision for the public body - may depend upon many considerations, including what is unknown and the impact of the thing that is predicted [52]. The point at which that determination is made, however, could inadvertently be moved back to the model-creators by the way that outputs are presented.

The extent to which algorithms 'may shift the style of decision-making towards specific rules and away from professional judgement and discretion' [7] (potentially in the process creating a challenge to the sovereignty of Parliament) is an area that requires further research within the practical environments in which algorithmic tools are implemented. If in practice we see the tone set within a public body that an algorithm's prediction must be followed, and thus the nature of the question to be answered changes, then I predict that, sooner rather than later, we will see a challenge based (partly) on a failure to consider the 'merits of the case' and a failure to consider other relevant factors. If on the other hand they can be designed to be a genuine aid to a public sector decision-maker's discretion, choices and professional judgement, without pushing the decision-maker towards a particular outcome, then there will be value. People are influenced by how information is 'framed.' [73] It is vital therefore that attention is paid to the human-algorithmic interface, together with the underlying organisational culture and processes to determine when a model's forecasts should be overriden, to minimise the risk of 'judgmental atrophy' leading to improper delegation to an algorithm. As Hartzog comments, 'Design affects our expectations about how things work and the context within which we are acting.' [74]

Improper delegation and fettering discretion – reframing for algorithm-assisted decision-making in the public sector

The third of my 're-framings' therefore considers the preservation of appropriate discretion in algorithm-assisted environments:

> The decision-making process, of which the algorithmic tool is part, must preserve the human discretion to assess 'un-thought of' relevant factors, and to assess whether the question or decision is the one for which the algorithm was designed. Algorithms should not be inserted into a process that requires the exercise of discretion by a public authority where the algorithm prevents that discretion; either because all of the factors relevant to the decision cannot be included, or required elements of the decision itself cannot be appropriately codified into, or by, the algorithm.

Conclusion

Cheney-Lippold, in his critique of our datafied culture, comments:

> 'We can think of a measurable type like 'at risk' as a hieroglyph, not a truth of identity but a priestly interpretation. It is not simply an officer who decides our fate in any given encounter with the police. Rather, it's an algorithmic interpretation of our own, datafied social networks that enacts police suspicion.' [75]

We might consider if this concern is anything new. Our eighteenth century clergymen show us that attempting to draw conclusions or make predictions from data is not a modern phenomenon. For our current society, however, it is surely the integration of (sometimes opaque) algorithmic conclusions and predictions into everyday life, powered by vast quantities of digital data, that creates both new opportunities and new challenges. Much of the concern around the growing ubiquity of algorithms in society can be boiled down to two questions: how do they work, and are the decisions made using algorithms fair? For centuries, English administrative law has been concerned with the fairness of state decisions. Its principles are already tech-agnostic. It has tackled issues of transparency and understanding, the relevance of 'inputs' and the protection of appropriate human discretion. For lawyers, scientists and public sector practitioners alike, old law – interpreted in a new context - can help guide our algorithmic-assisted future.

References

1. Reverend Mr. Barrow. 1735. An account of the births and burials with the number of the inhabitants at Stoke-Damerell in the county of Devon. Communicated by the Reverend Mr. Barrow. *Phil. Trans. R. Soc. Lond.* **39**. 171-172.
2. William Morgan, communicated by the Rev. Richard Price. 1789. 'On the method of determining, from the real probabilities of life, the value of a contingent reversion in which three lives are involved in the survivorship' *Phil. Trans. R. Soc. Lond.* **79**. 40-54; Richard Price was the friend of Thomas Bayes responsible for reading Bayes's rule to the Royal Society after his death: Rev. Mr. Bayes, F.R.S. 1763. 'An essay towards solving a problem in the doctrine of chances. By the late Rev. Mr. Bayes, F.R.S. communicated by Mr. Price, in a letter to John Canton, A.M.F.R.S' *Phil. Trans.* **53**. 370-418.
3. Robert Hassan. 2008. *The Information Society* Polity Press.
4. William M. Grove and Paul E. Meehl 1996. Comparative efficiency of informal (subjective, impressionistic) and formal (mechanical, algorithmic) prediction procedures: The clinical–statistical controversy. *Psychology, Public Policy, and Law,* ***2*(2)**, 293-323.
5. Commission Nationale Informatique & Libertés. 2017. *How can humans keep the upper hand? The ethical matters raised by algorithms and artificial intelligence. Report on the public debate led by the French Data Protection Authority (CNIL) as part of the ethical discussion assignment set by the Digital Republic Bill*. https://www.cnil.fr/sites/default/files/atoms/files/cnil_rapport_ai_gb_web.pdf.
6. Policy paper: Government Transformation Strategy 2017 to 2020, 9 February 2017 https://www.gov.uk/government/publications/government-transformation-strategy-2017-to-2020.
7. Andrew Le Sueur. 2016. 'Robot Government: Automated Decision-Making and its Implications for Parliament' in Alexander Horne and Andrew Le Sueur (ed.) *Parliament: Legislation and Accountability* Hart Publishing.

8. For a summary of four explanation styles 'input influence', 'sensitivity', 'case-based' and 'demographic' see Reuben Binns, Max Van Kleek, Michael Veale, Ulrik Lyngs, Jun Zhao and Nigel Shadbolt. 2018. 'It's Reducing a Human Being to a Percentage'; Perceptions of Justice in Algorithmic Decisions. *ACM Conference on Human Factors in Computing Systems (CHI'18), April 21–26, Montreal, Canada*. doi: 10.1145/3173574.3173951.
9. Julia Angwin, Jeff Larson, Surya Mattu and Lauren Kirchner. 2016. Machine Bias. *ProPublica* https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing; Julia Dressel and Hany Farid. 2018. The accuracy, fairness and limits of predicting recidivism *Sci. Adv.* **4 (1)** DOI: 10.1126/sciadv.aao5580.
10. Marion Oswald, Jamie Grace, Sheena Urwin and Geoffrey Barnes. 2018. Algorithmic Risk Assessment Policing Models: Lessons from the Durham HART Model and 'Experimental' Proportionality. *Information & Communications Technology Law,* **27(2)**, 223-250. https://www.tandfonline.com/doi/abs/10.1080/13600834.2018.1458455.
11. Dan Hurley. 2018. Can an algorithm tell when kids are in danger? *New York Times* https://www.nytimes.com/2018/01/02/magazine/can-an-algorithm-tell-when-kids-are-in-danger.html.
12. Rhema Vaithianathan. 2017. Five Lessons for Implementing Predictive Analytics in Child Welfare. *The Chronicle of Social Change* https://chronicleofsocialchange.org/opinion/five-lessons-implementing-predictive-analytics-child-welfare.
13. Virginia Eubanks. 2018. A child abuse prediction model fails poor families. *Wired*. https://www.wired.com/story/excerpt-from-automating-inequality/.
14. Jenna Burrell. 2016. How the machine 'thinks': Understanding opacity in machine learning algorithms. *Big Data & Society*. **January – June 2016**.
15. Geoffrey Barnes and Jordan M. Hyatt. 2012. Classifying Adult Probationers by Forecasting Future Offending Final Technical Report https://www.ncjrs.gov/pdffiles1/nij/grants/238082.pdf.
16. Riccardo Guidotti, Anna Monreale, Franco Turini, Dino Pedreschi and Fosca Giannotti. 2018. A Survey of Methods For Explaining Black Box Methods. *arXiv:1802.01933v1 [cs.CY]*.
17. Stanton and Prescott. 2018 *Public Law* Oxford: OUP.
18. *R v Secretary of State for the Home Department ex p. Fayed* [1997] 1 All ER 763.
19. Wade and Forsyth. 2014 *Administrative Law* 10th ed. Oxford: OUP.
20. *Clarke Holmes Ltd v Secretary of State for the Environment* (1993) 66 P & CR 263, 271-272.
21. *South Buckinghamshire District Council v Porter (No 2)* [2004] 1 WLR 1953, para 36.
22. *Dover District Council (Appellant) v CPRE Kent (Respondent) CPRE Kent (Respondent) v China Gateway International Limited (Appellant)* [2017] UKSC 79, para 41.
23. *R v Secretary of State for the Home Department ex parte Doody* [1994] 1 AC 531, 19 (Lord Mustill).

24. Mireille Hildebrandt. 2017. Privacy As Protection of the Incomputable Self: Agonistic Machine Learning. Available at SSRN: https://ssrn.com/abstract=3081776 or http://dx.doi.org/10.2139/ssrn.3081776 .
25. Article 29 Data Protection Working Party Guidelines on Automated individual decision-making and Profiling (WP251rev.01)
26. It is possible to envisage that the 'Counterfactual' method identifying key factors and weightings might be adapted for this purpose: Sandra Wachter, Brent Mittelstadt and Chris Russell. 2017. Counterfactual Explanations without Opening the Black Box: Automated Decisions and the GDPR. *arXiv:1711.00399v2 [cs.AI].*
27. General Medical Council 'Consent: patients and doctors making decisions together' 2 June 2008 https://www.gmc-uk.org/ethical-guidance/ethical-guidance-for-doctors/consent
28. Courts and Tribunal Judiciary: Judicial Accountability and Independence https://www.judiciary.uk/about-the-judiciary/the-judiciary-the-government-and-the-constitution/jud-acc-ind/.
29. Shai Danziger, Jonathan Levav and Liora Avnaim-Pesso. 2011. 'Extraneous factors in judicial decisions' *PNAS* 108(17) 6889-6892. Using Data Science in Policy *A report by the Behavioural Insights Team*. December 14, 2017. http://www.behaviouralinsights.co.uk/publications/using-data-science-in-policy/.
30. Kren Weinshall-Margel and John Shapard. 2011. 'Overlooked factors in the analysis of parole decisions' *PNAS* **108 (42)** E833.
31. Andreas Glöckner. 2016. 'The irrational hungry judge effect revisited: Simulations reveal that the magnitude of the effect is overestimated' *Judgment and Decision Making* **11(6)** 601-610.
32. Frank Pasquale and Glyn Cashwell. 2018. 'Prediction, persuasion, and the jurisprudence of behaviourism' *University of Toronto Law Journal* **68(1)** 63-81.
33. Mireille Hildebrandt. 2016. New Animism in Policing: Re-animating the Rule of Law? In *The Sage Handbook of Global Policing* (ed. Bradford, Jauregui, Loader and Steinberg) Sage.
34. Nick Polson and James Scott. 2018. *AIQ: How artificial intelligence works and how we can harness its power for a better world* Bantam Press.
35. David Martens and Foster Provost. 2014. 'Explaining Data-Driven Document Classifications' *MIS Quarterly* **38(1)** 73-99.
36. HM Treasury. 2015. *The Aqua Book: guidance on producing quality analysis for government* https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/416478/aqua_book_final_web.pdf.
37. Areas include academic success, business bankruptcy and parole violation in Dawes, R.M., Faust, D., & Meehl, P.E. 1993. Statistical prediction versus clinical prediction: Improving what works. In G. Keren & C. Lewis (Eds.) *A handbook for data analysis in the behavioral sciences: Methodological issues* (pp. 351 – 367). Hillsdale, NJ: Lawrence Erlbaum.
38. Domestic violence risk assessments in Richard A. Berk, Susan B. Sorenson and Geoffrey Barnes. 2016. 'Forecasting Domestic Violence: A Machine Learning Approach to Help Inform Arraignment Decisions' *Journal of Empirical Legal Studies* **13 (1)** 94-115.

39. Richard A. Berk and Justin Bleich. 2013. 'Statistical Procedures for Forecasting Criminal Behaviour: A Comparative Assessment' *Criminology & Public Policy* **12(3)** 513-544.
40. Max Van Kleek, William Seymour, Michael Veale, Reuben Binns, Nigel Shadbolt. 2018. 'The need for sensemaking in networked privacy and algorithmic responsibility' Sensemaking Workshop: CHI 2018, Montréal, Canada https://hip.cat/papers/sensemaking-networked-privacy.pdf.
41. Geoffrey C. Barnes interviewed on BBC Click 'Data Detectives' 5 May 2018 https://www.bbc.co.uk/iplayer/episode/b0b268rx/click-data-detectives#.
42. *R (DSD & Anor) v The Parole Board of England and Wales* [2018] EWHC 694 (Admin).
43. Michael Veale, Max Van Kleek and Reuben Binns. 2018. Fairness and Accountability Design Needs for Algorithmic Support in High-Stakes Public Sector Decision-Making. *arXiv:1802.01029 [cs.CY]*.
44. Angele Christin. 2017. 'Algorithms in practice: Comparing web journalism and criminal justice' *Big Data & Society* **July – Dec** 1-14.
45. *R v St Pancras Vestry* (1890) 24 QBD 371, 375.
46. *R v Home Secretary ex p. Venables* [1998] AC 407.
47. *Short v. Poole Cpn.* [1926] Ch. 66.
48. Rhema Vaithianathan et al. 2017. 'Developing Predictive Models to Support Child Maltreatment Hotline Decisions: Allegheny County Methodology and Implementation'. https://www.alleghenycountyanalytics.us/wp-content/uploads/2018/02/DevelopingPredictiveRiskModels-package_011618.pdf.
49. Federico Cabitza, Davide Ciucci and Raffaele Rasoini. 2018. 'A giant with feet of clay: on the validity of data that feed machine learning in medicine' *arXiv:1706.06838 [cs.LG]*.
50. Marco Tulio Ribeiro, Sameer Singh and Carlos Guestrin. 2016. "Why Should I Trust You?" Explaining the Predictions of Any Classifier. *arXiv:1602.04938v3 [cs.LG].*
51. Melissa Hamilton. 2015. 'Adventures in risk: Predicting Violent and Sexual Recidivism in Sentencing Law' *Arizona State Law Journal*, **47 (1)** 1-62.
52. Michael Blastland and David Spiegelhalter. 2013 *The Norm Chronicles* Profile Books.
53. *State of Wisconsin v Eric L. Loomis*, 2016 WI 68.
54. House of Commons Science and Technology Committee 'Algorithms in decision-making' Fourth Report of Session 2017 – 19 HC 351 23 May 2018.
55. Judea Pearl and Dana Mackenzie. 2018. *The Book of Why* Allen Lane.
56. Jake M. Hofman, Amit Sharma and Duncan J. Watts. 2017. Prediction and explanation in social systems. *Science*.
57. See for instance Julia Dressel and Hany Farid. 2018. The accuracy, fairness and limits of predicting recidivism *Sci. Adv.* **4 (1)** DOI: 10.1126/sciadv.aao5580.
58. See Jamie Grace. 2015. 'Clare's Law, or the national Domestic Violence Disclosure Scheme : the contested legalities of criminality information sharing' *The Journal of Criminal Law* **79** (1) 36-45.
59. Geoff Mulgan. 2018. *Big Mind: How Collective Intelligence Can Change Our World* Princeton University Press.

60. *Barnard v National Dock Labour Board* [1953] 2 QB 18.
61. *Ellis v Dubowski* [1921] 3 KB 621.
62. *R v Home Secretary ex p. Walsh* [1992] COD 240.
63. *British Oxygen Co Ltd v Minister of Technology* [1971] AC 610, 625.
64. *Merchandise Transport Ltd v British Transport Commission* [1962] 2 QB 173.
65. Mireille Hildebrandt. 2017. Law As Computation in the Era of Artificial Legal Intelligence. Speaking Law to the Power of Statistics. Available at SSRN: https://ssrn.com/abstract=2983045.
66. Karen Yeung. 2017. 'Hypernudge': Big Data as a Mode of Regulation by Design. *Information, Communication & Society* **20**(1).
67. Garry Kasparov. 2017 *Deep Thinking: Where Machine Intelligence Ends And Human Creativity Begins* John Murray.
68. Jon Kleinberg, Himabindu Lakkaraju, Jure Leskovec, Jens Ludwig and Sendhil Mullainathan. 2017. Human Decisions and Machine Predictions *NBER Working Paper No. 23180.* National Bureau of Economic Research. https://www.cs.cornell.edu/home/kleinber/w23180.pdf.
69. Min Kyung Lee. 2018. 'Understanding perception of algorithmic decisions: Fairness, trust, and emotion in response to algorithmic management' *Big Data & Society* **Jan – June** 1-16.
70. https://www.documentcloud.org/documents/2702103-Sample-Risk-Assessment-COMPAS-CORE.html
71. Authorised Professional Practice: Risk Principles https://www.app.college.police.uk/app-content/risk-2/risk/.
72. Police and Criminal Evidence Act 1984 c.60 s38(1)(a).
73. Cass R. Sunstein. 2013 *Simpler: The Future of Government* New York: Simon & Schuster.
74. Woodrow Hartzog. 2018. *Privacy's Blueprint: The Battle to Control the Design of New Technologies* Harvard University Press.
75. John Cheney-Lippold. 2017 *We are Data: Algorithms and the making of our digital selves* New York University Press.